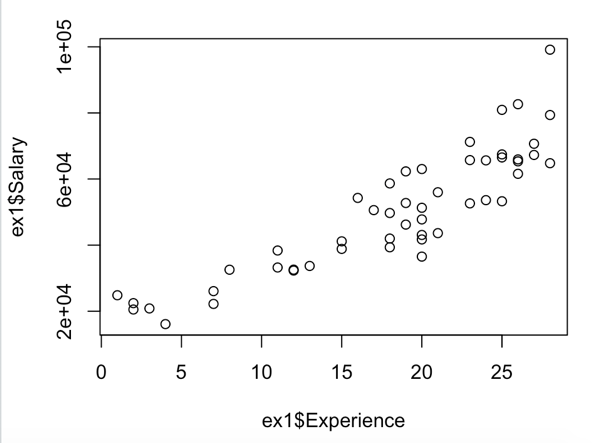
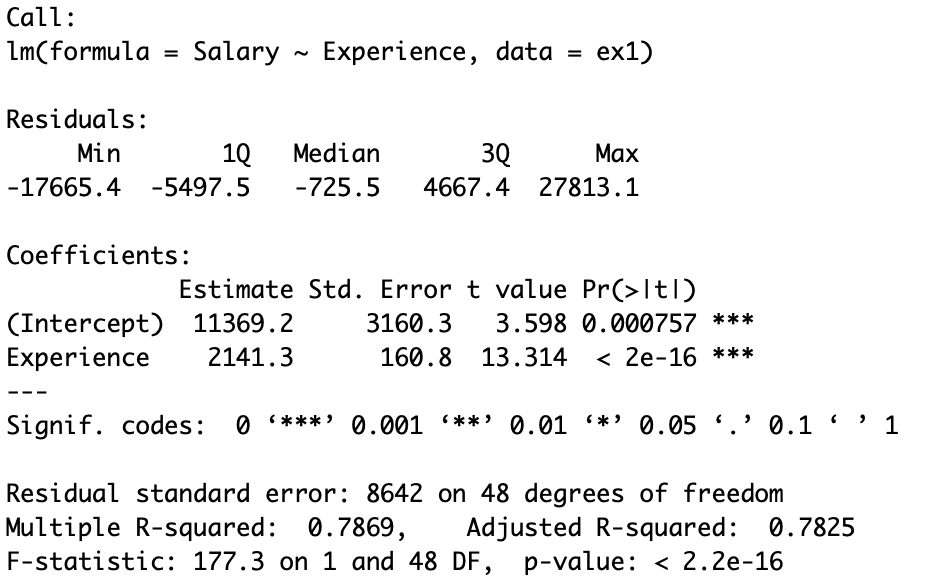
BSAN 450 Assignment 1

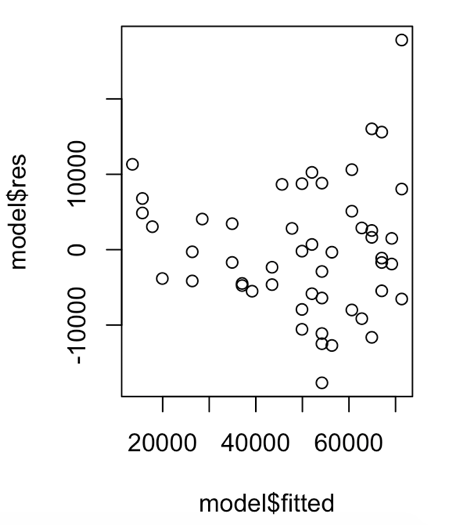
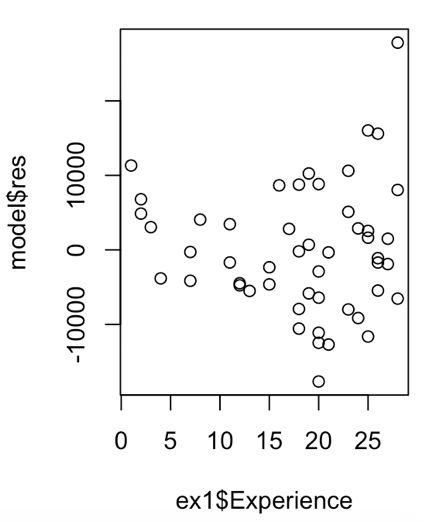
1) In this example there are 2 variables: Salary = the salary of the managers in a company and Experience = the years of experience for the manager. The objective is to fit a model for the Salary in terms of the Experience. The data is in a file named “ManSalary.csv”. After you set the working directory in which this file is stored on your computer, read the data into R Studio.

a) Plot a scatter plot of the Salary on the vertical axis and Experience on the horizontal axis. Comment on this plot. Is there a relationship between these two variables? How would you describe this relationship?

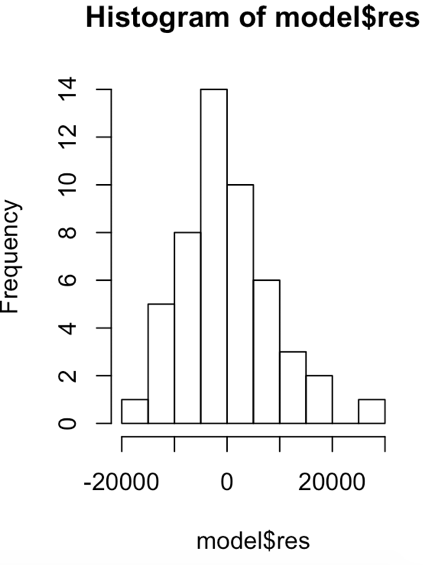
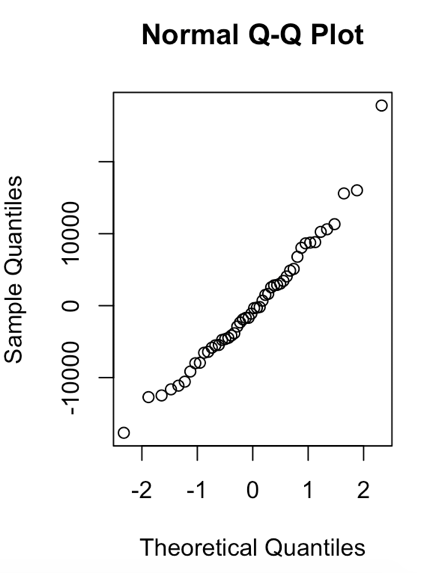
It looks like there is a positive linear relationship between the experience a worker has and the salary that they make.

b) Estimate a simple linear model with the Y variable equal to Salary and the X variable equal to Experience. Print out a summary of this model. 

c) Perform the diagnostic checks of the residuals for the model you estimated in part b). For each plot is there any indication of a problem with the model?



The residual vs fitted values and the residual vs experience plots look fine. The points are evenly spread, there appears to be no apparent relationship, and the mean is about 0.



The histogram has the shape of a bell curve which is expected.

The normal plot looks as expected in that it is a straight line.

Performing the Shapiro-Wilk normality test results in a p-value of 0.3219 which is less than .5 therefore the hypothesis that the data follows a normal distribution is rejected. This is a problem for the model estimated in part b.

d) Test the hypothesis that the slope of the regression line equal 0 versus the alternative hypothesis that the slope of the regression line is not equal to 0. What can you conclude from this test?

Given that the p-value for this data is < 2.2e-16, the null hypothesis can be rejected. We can conclude that experience does affect salary.

e) What is the value of R-squared for this example? What does this value tell you about this model?

The R-squared value is 0.7869 or 78.69%. This tells us that 78.69% of the variation in salary is explained by experience.

f) Find a 95% confidence interval for the parameters .

The confidence interval for is (5015.095, 17723.352).

The confidence interval for is (1817.927, 2464.693).

The following R commands will produce the plots and the output. You should make sure you know what these commands are doing so that you could do the same things with another set of data.

sal=read.csv("ManSalary.csv")

plot(sal$Salary~sal$Experience)

fit=lm(Salary~Experience, data=sal)

summary(fit)

plot(fit$res~fit$fitted)

plot(fit$res~sal$Experience)

hist(fit$res)

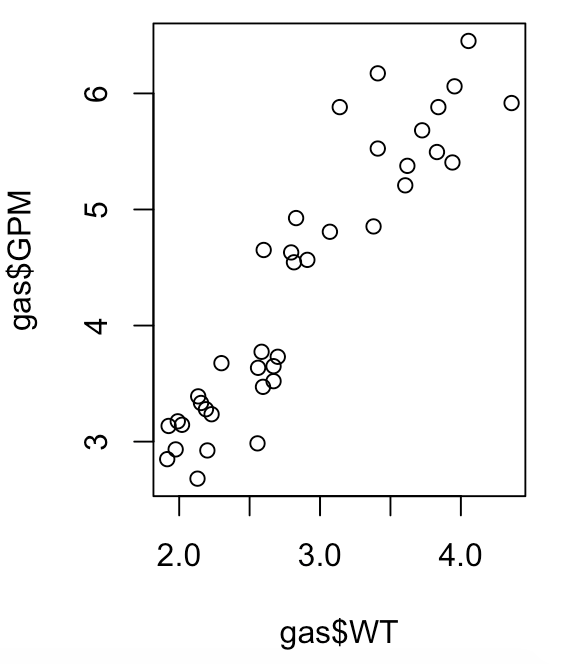
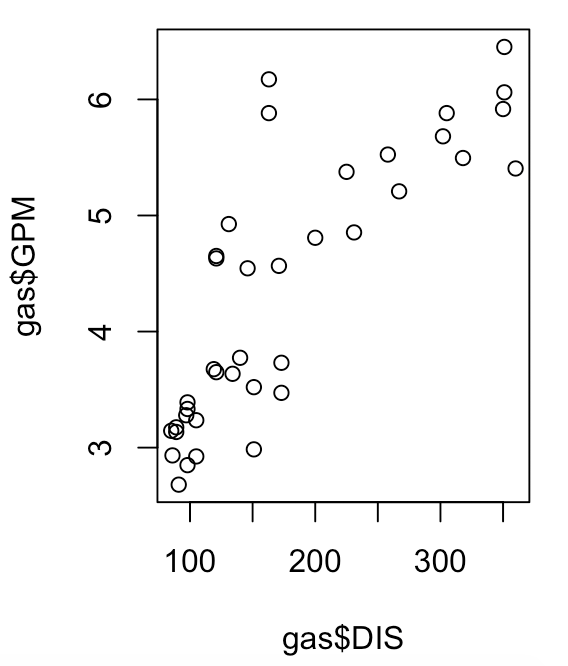
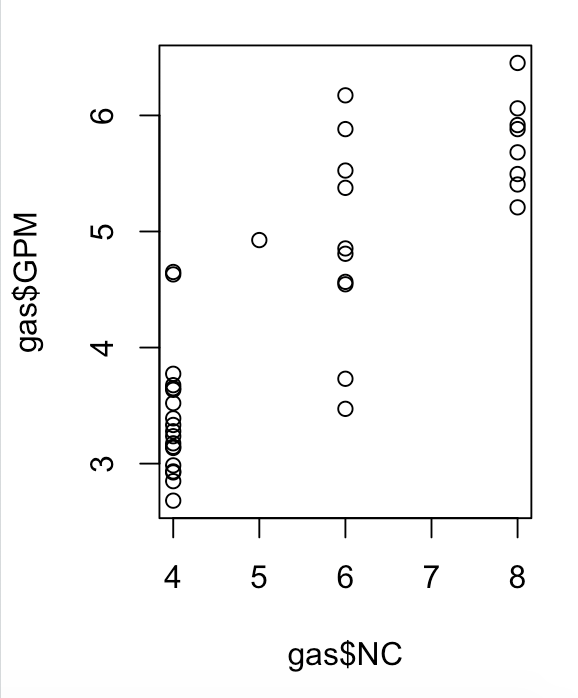
qqnorm(fit$res)

shapiro.test(fit$res)

confint(fit)

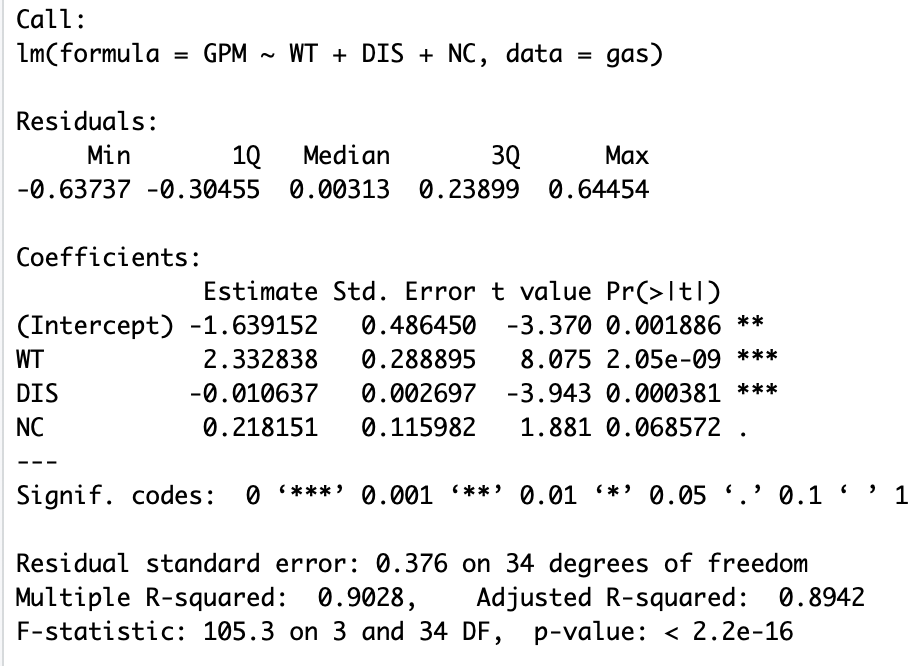
2) In this example we are interested in predicting the gas consumption of an automobile based on characteristics of the automobile. The data consists of 38 cars with measurements on fuel efficiency, weight of the car, engine displacement, and number of cylinders. The data is in the file gasconsumption.csv. The names of the variables are GPM = gallons used per 100 miles, WT = weight of the car in 1000 pounds, DIS = displacement of the engine in cubic inches, and NC = the number of cylinders in the engine. You are to develop a model for GPM based on the characteristics of the car, WT, DIS, and NC.

a) Read the data into R Studio and plot scatter plots of GPM versus all the X variables. Comment on these plots. Is there a relationship between the GPM and the possible X variables? How would you describe the relationship?

Gallons vs the weight of the car and gallons vs the displacement of the engine appear to have positive linear relationships. Gallon vs weight appears to have a stronger relationship of the two. Gallons vs the number of cylinders appears to have no noticeable relationship.

b) Estimate a multiple regression model with the Y variable equal to GPM and the X variables equal to WT, DIS, and NC. Print out a summary of this model.



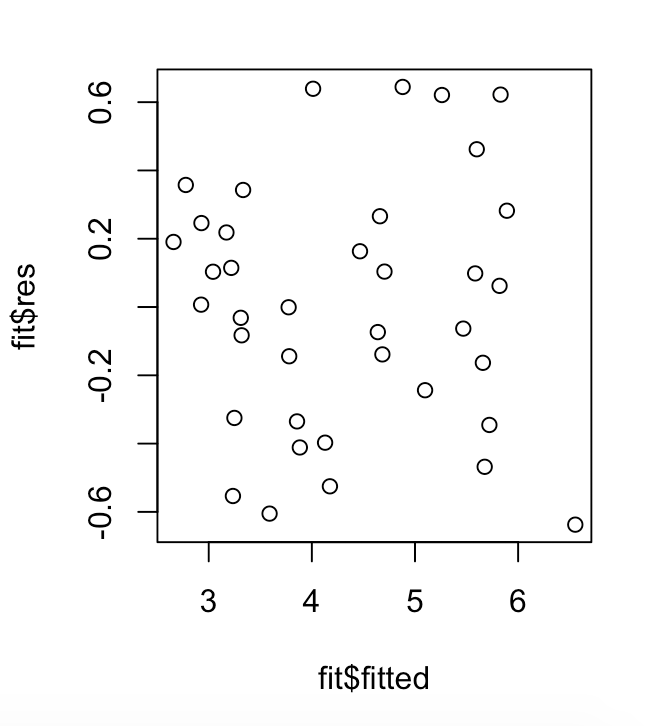
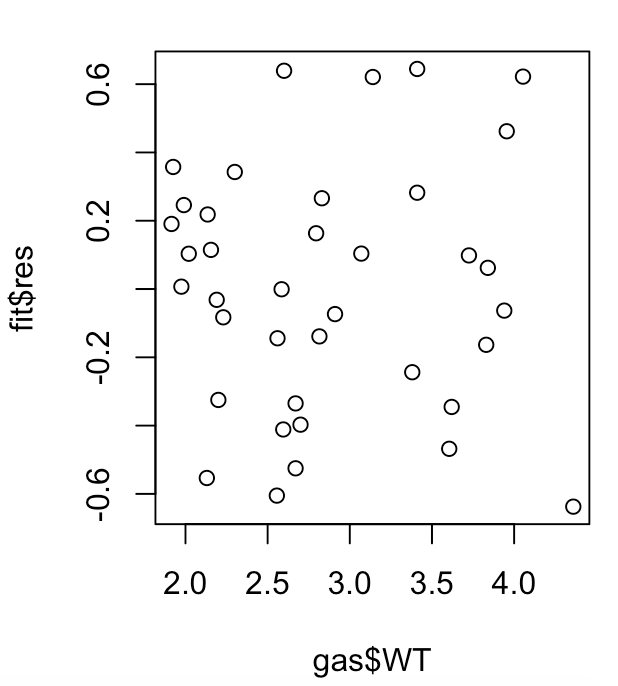
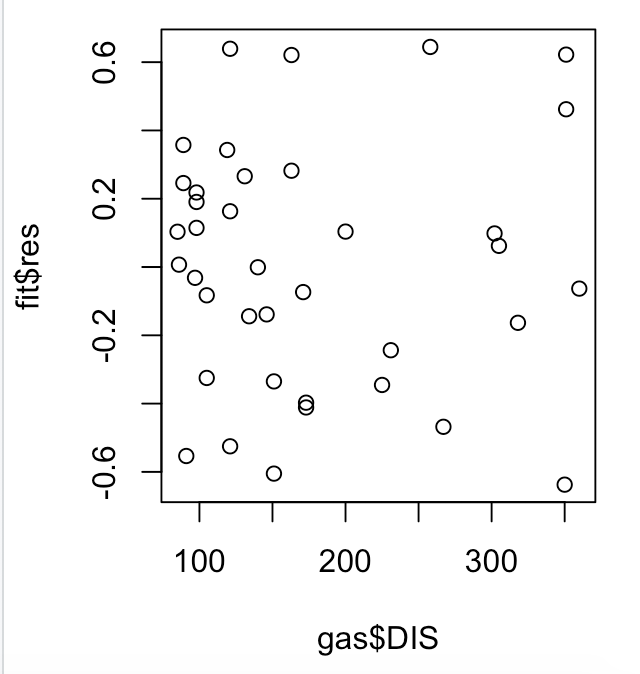
c) Use the output for this model to test the null hypothesis: = 0 versus the alternative hypothesis: not all ’s are equal to 0. What does your result mean?

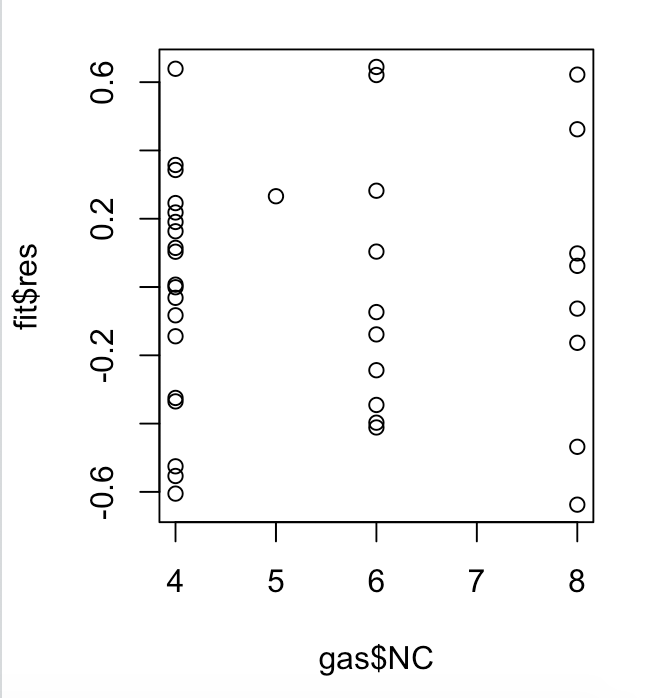
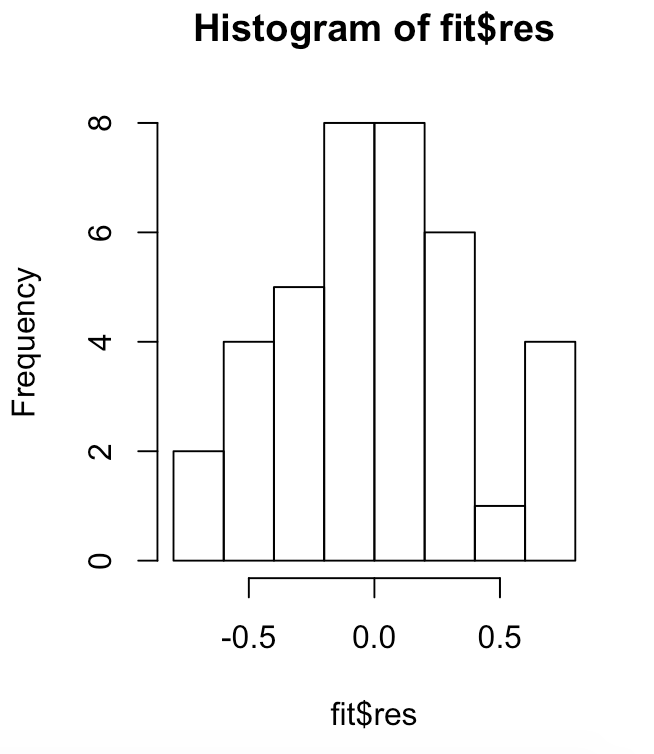
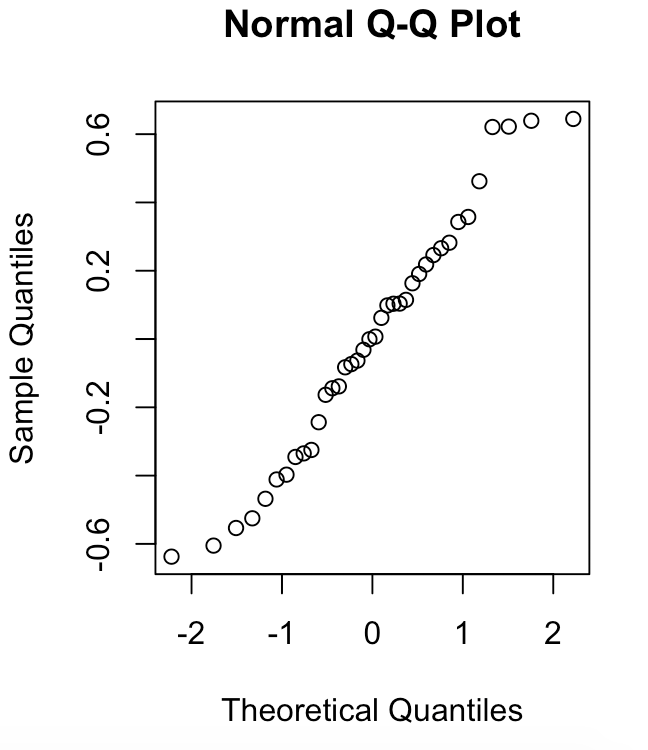
The null hypothesis is rejected because even though the p-value for the number of cylinders is .068572 which is greater than 0.05 the p-value for the other two independent variables are less than 0.05. This means that the number of cylinders does not impact the GMP, but the weight of the car and the displacement of the engine do.

d) Based upon the output, are all of the 3 variables WT, DIS, and NC needed in the model? Justify your answer.

No, the p-value for the number of cylinders variable is above 0.05 which indicates no significant impact. However, the weight and the displacement of the engine have p-values less than 0.05 which indicates an impact.

e) Perform the usual diagnostic checks for this model. Based upon your analysis is there any problems with this model?

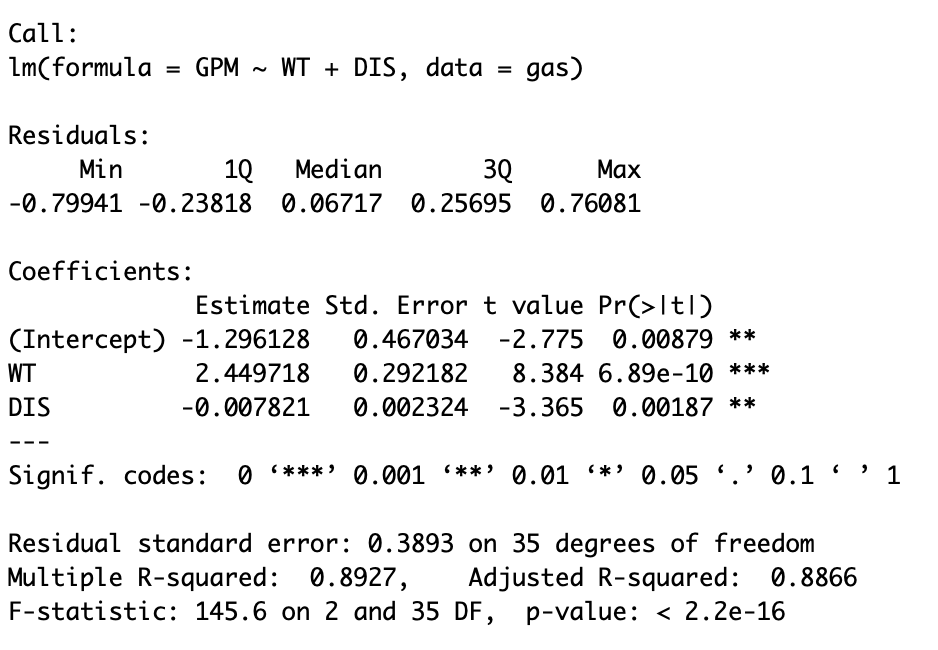
The residual vs fitted values, the residual vs WT, the residual vs DIS, and the residual vs NC plots look fine. There appears to be no apparent relationship, and the mean is about 0. The data points in the residual vs NC are slightly more clustered with 4 cylinders than for 6 or 8. The data points for the residual vs WT and the residual vs DIS are well dispersed.

The histogram has the shape of a bell curve which as expected. The normal plot looks as expected in that it is approximately a straight line.

Performing the Shapiro-Wilk normality test results in a p-value of 0.3816 which is less than .5 therefore the hypothesis that the data follows a normal distribution is rejected. This is a problem for the model estimated in part b.

f) Is there anything that you would do to change the model? If yes, then make the appropriate changes and estimate a new model.

I would change the model by excluding the number of cylinders (NC) as it does not have a strong impact on the GPM.



The following R commands will produce the plots and the output, except for possibly part f). You should make sure you know what these commands are doing so that you could do the same things with another set of data.

gas=read.csv("gasconsumption.csv")

plot(gas$GPM~gas$WT)

plot(gas$GPM~gas$DIS)

plot(gas$GPM~gas$NC)

fit=lm(GPM~WT+DIS+NC, data=gas)

summary(fit)

plot(fit$res~fit$fitted)

plot(fit$res~gas$WT)

plot(fit$res~gas$DIS)

plot(fit$res~gas$NC)

hist(fit$res)

qqnorm(fit$res)

shapiro.test(fit$res)